tvst

Article

A Deep Learning Algorithm for Classifying Diabetic Retinopathy Using Optical Coherence Tomography Angiography

Gahyung Ryu[1,2,*], **Kyungmin Lee**[3,*], **Donggeun Park**[1], **Inhye Kim**[1], **Sang Hyun Park**[3], **and Min Sagong**[1,4]

[1] Department of Ophthalmology, Yeungnam University College of Medicine, Daegu, South Korea
[2] Nune Eye Hospital, Daegu, South Korea
[3] Department of Robotics Engineering, DGIST, Daegu, South Korea
[4] Yeungnam Eye Center, Yeungnam University Hospital, Daegu, South Korea

Correspondence: Min Sagong, Department of Ophthalmology, Yeungnam University College of Medicine, #170 Hyunchungro, Nam-gu, Daegu 42415, South Korea. e-mail: msagong@yu.ac.kr
Sang Hyun Park, Department of Robotic Engineering, DGIST, #333, Techno jungang-daero, Dalseong-gun, Daegu, South Korea. e-mail: shpark13135@dgist.ac.kr

Received: October 7, 2021
Accepted: February 7, 2022
Published: February 28, 2022

Keywords: deep convolutional neural network; diabetic retinopathy; machine learning; optical coherence tomography angiography; radiomics

Citation: Ryu G, Lee K, Park D, Kim I, Park SH, Sagong M. A deep learning algorithm for classifying diabetic retinopathy using optical coherence tomography angiography. Transl Vis Sci Technol. 2022;11(2):39, https://doi.org/10.1167/tvst.11.2.39

Purpose: To develop an automated diabetic retinopathy (DR) staging system using optical coherence tomography angiography (OCTA) images with a convolutional neural network (CNN) and to verify the feasibility of the system.

Methods: In this retrospective cross-sectional study, a total of 918 data sets of 3×3 mm^2 OCTA images and 917 data sets of 6×6 mm^2 OCTA images were obtained from 1118 eyes. A deep CNN and four traditional machine learning models were trained with annotations made by a retinal specialist based on ultra-widefield fluorescein angiography. Separately, the same images of the test data sets were independently graded by two human experts. The results of the CNN algorithm were compared with those of traditional machine learning–based classifiers and human experts.

Results: The proposed CNN achieved an accuracy of 0.728, a sensitivity of 0.675, a specificity of 0.944, an F1 score of 0.683, and a quadratic weighted κ of 0.908 for a six-level staging task, which were far superior to the results of traditional machine learning methods or human experts. The CNN algorithm showed a better performance using 6×6 mm^2 rather than 3×3 mm^2 sized OCTA images and using combined data rather than a separate OCTA layer alone.

Conclusions: CNN-based classification using OCTA images can provide reliable assistance to clinicians for DR classification.

Translational Relevance: This CNN algorithm can guide the clinical decision for invasive angiography or referrals to ophthalmology specialists, helping to create more efficient diagnostic workflow in primary care settings.

Introduction

Diabetic retinopathy (DR), a leading cause of blindness worldwide, can be delayed or prevented through appropriate treatment.[1–4] Because the success of such treatment depends on timely interventions, most guidelines recommend regular DR screening for diabetic patients.[5,6] However, existing screening methods can miss a substantial fraction of DR cases, leading to preventable vision loss because DR grading based on traditional fundus images is often subjective depending on expert clinical interpretation.[7,8] In addition, the requirement for highly trained ophthalmologists is an expensive and time-consuming process, making it infeasible to screen all diabetes patients for DR.[9]

Deep learning application for automated retinal image analysis has recently demonstrated specialist-level accuracy in the diagnosis of DR severity, substantially aiding access to DR screening and improving the diagnostic accuracy.[10–21] However, these studies have not addressed a fundamental weakness of their own: the uncertain accuracy of data annotation. Vascular alterations caused by diabetes are widely distributed, and up to 50% of DR lesions are known to be located outside seven-standard defined retinal fields.[22,23] Multiple independent groups have reported that peripheral retinal lesions outside standard fields suggest a more severe DR grade in 9% to 19% of eyes.[22,24,25] Although Early Treatment Diabetic Retinopathy Study classification is still the major staging system for DR, a rigorous evaluation of the retina using ultra-widefield (UWF) fluorescein angiography (FA) can be a more accurate method of assessing DR severity, without relying on specific features (microaneurysms or bleeding) provided by fundus photographs with a lower resolution and limited field of view. However, invasive FA is unsuitable as a screening tool or for use in frequent longitudinal assessments because it carries the risk of serious adverse reactions.[26]

Although optical coherence tomography angiography (OCTA) provides a limited field of view compared to UWF FA, it has a marked advantage of being a noninvasive, rapid, and simple approach providing detailed three-dimensional information of the retinal vascular network.[27,28] Accordingly, whether OCTA can substitute UWF FA has been investigated, and several studies have reported that OCTA imaging is useful even for estimating peripheral capillary perfusion.[29–33] However, they used only predetermined quantitative features of an OCTA analysis, relying only on empirically selected biomarkers. To identify even the subtle microvascular changes caused by diabetes across the retina, it is necessary to use a much richer feature space, latent within all OCTA data.

In this study, we presented an end-to-end deep convolutional neural network (CNN)–based method for classifying DR severity automatically from OCTA images. The approach was trained and tested with annotations based on UWF FA. The feasibility of the proposed model was confirmed through a quantitative comparison of the model performance against four different machine learning–based classifiers that use handcrafted features extracted from OCTA images and human graders.

Methods

This cross-sectional study was approved by the Institutional Review Board (IRB) of Yeungnam University Medical Center (IRB number, 2020-09-079) and conducted in accord with the Declaration of Helsinki. The requirement for written consent was waived by the IRB because of the retrospective nature of the study. Data were collected between January 2018 and July 2020.

For normal eyes, patients who had visited the hospital for visual floater and ocular discomfort and who had undergone detailed examination, including OCTA (Optovue RTVue XR AVANTI; Optovue, Inc., Fremont, CA, USA), but had no systemic disease or ocular disease were included. For diabetic eyes, patients who had previously been diagnosed with diabetes mellitus (DM) and undergone comprehensive ophthalmic examinations, including UWF FA (Optos California; Optos plc, Dunfermline, UK) and OCTA, were included. Because this was retrospective chart review study, indication of invasive angiography was not related to the protocol of this study. UWF FA was performed limitedly after explaining possible side effects if the patient desired a full-examination despite absence of DR. Exclusion criteria included the presence of glaucoma or retinal disorders affecting retinal capillary changes other than DR. Eyes with macular edema (defined as a retinal thickening of at least 315 µm and/or intra- or subretinal fluid seen on the optical coherence tomography [OCT] B-scan) and media opacity precluding imaging were excluded because this can obscure retinal microvasculature on OCTA. Images with a low signal strength ($\leq$6), excessive motion, or projection artifacts were also excluded. OCTA images were obtained as volume scans of 3 × 3 mm^2 and 6 × 6 mm^2 in size centered on the macula, and images of the superficial capillary plexus (SCP), deep capillary plexus (DCP), and full-thickness retinal slab were used for analysis. The ground truth labels for both training and testing were grades previously assigned to UWF FA images by an expert human grader (MS) based on the International Clinical Diabetic Retinopathy Severity Scale, which was adapted by means of extending the grading quadrants to the periphery of the entire image while maintaining the original grading nomenclature for simplicity.[34]

The overall study design for the classification of the DR is shown in Figure 1. OCTA images and demographic data, including age and sex, were used as inputs. Each of the six performance metrics obtained from the CNN classifier, machine learning classifiers, and experts was evaluated and compared.

Development of CNN-Based Classifier

The overall structure of the proposed CNN-based end-to-end classifier for DR classification is shown

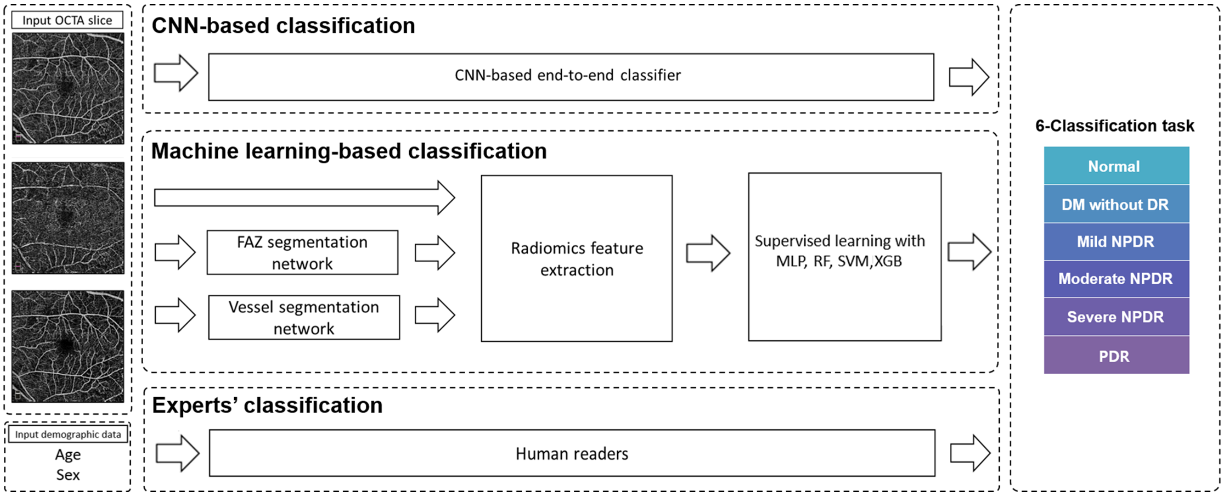

Figure 1. Overall study design for classification of DR.

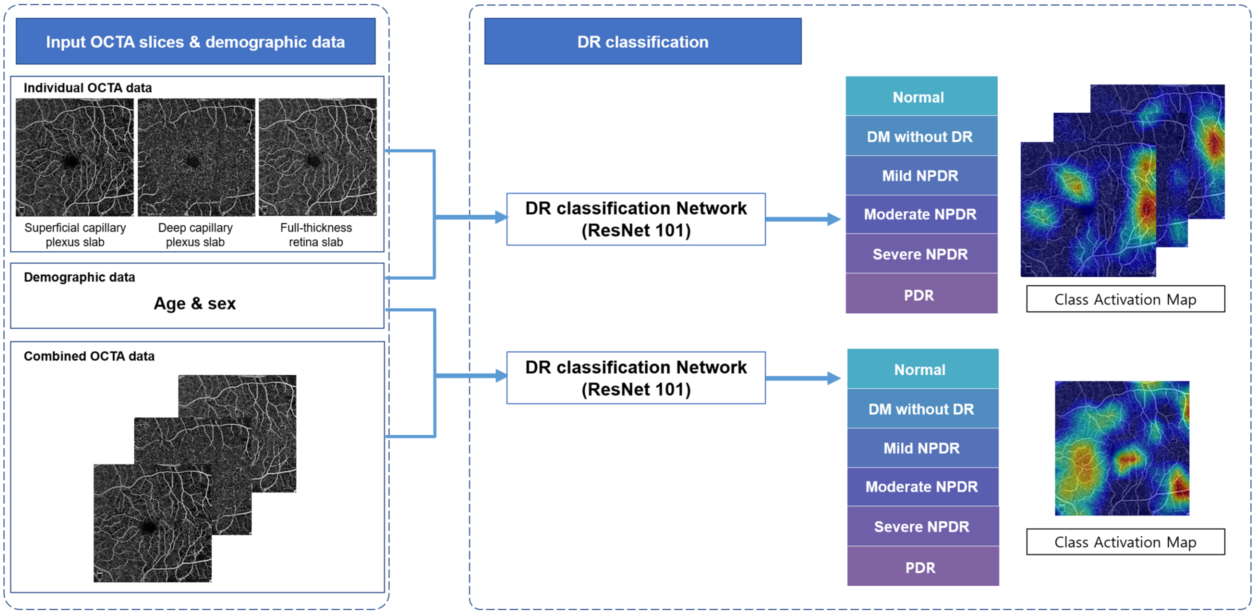

Figure 2. CNN-based classifier for DR. OCTA images and demographic data are passed through an end-to-end CNN-based classifier to derive the DR class prediction and class activation maps, which visualize the regions significantly affecting the tasks.

in Figure 2. SCP, DCP, and full-thickness retina OCTA images were concatenated and used as inputs to the CNN. Using the ResNet 101 model, images were passed through residual blocks with 101 layers. A summation of the input and output feature maps was repeatedly applied from the convolution layers, each with a batch normalization, rectified linear unit (ReLU) activation functions, and max pooling.[35] After passing through the residual blocks, each feature map was averaged in the global average pooling (GAP) layer. Using the averaged feature maps and demographic data, the probability of each stage was obtained through a fully connected layer with a softmax function. Class activation maps (CAMs) were derived from the GAP layer by summating the feature maps with the weights from the last layer to visualize regions that show a high correlation with the task of interest.[36] Notably, parameters of the network were initially transferred from the pretrained parameters of the ImageNet data set, excluding the first and last layer parameters.[37] Subsequently, all parameters were retrained using our OCTA data set, which was optimized based on the cross-entropy loss with an Adam optimizer and a learning rate of 0.0001.[38]

Development of Machine Learning–Based Classifier

The machine learning–based classifier consisted of three stages: segmentation, feature extraction, and classification (Fig. 3).

In the segmentation step, U-Net[39] was used to segment the blood vessels and foveal avascular zone (FAZ) from the OCTA images. The combined data from each layer of the OCTA image were used as input, given that prior machine learning–based studies demonstrated best DR classification results when local features from the combined data were used, that is, both SCP and DCP.[11,20,21,40] In the U-Net model, a contracting path extracts high-level features from the input images by repeatedly using convolution layers, batch normalization, ReLU activation function, and max pooling, whereas the expanding path generates a segmentation map of the same size as the input image by repeatedly using upsampling, convolution layers, batch normalization, and ReLU activation functions on the extracted high-level features. In the expanding path, intermediate feature maps of the contracting path were concatenated with the feature maps of the previous expanding path and used as input in the next expanding path. The parameters in the network were optimized using the Adam optimizer with a dice similarity coefficient loss and a learning rate of 0.0001.[38]

In the feature extraction stage, four local features (blood vessel density, skeletal vessel density, fractal dimension, and FAZ area) were extracted from the segmented OCTA images. In addition, 9 shape features (major and minor axis lengths, maximum diameter, elongation, sphericity, perimeter, perimeter surface ratio, number of meshes, and number of pixels), 18 intensity features (i.e., mean, median, standard deviation, and entropy of the intensity values), and 75 texture features were extracted from an image and its segmentation masks using the Pyradiomics toolbox.[41] The intensity and texture features were extracted from the original input image as well as the image filtered by Laplacian or Gaussian filters with different sigma values (i.e., 1, 2, and 3). Thus, $9 + (18 + 75) \times 4 = 381$ features were extracted for each segmentation mask. The features were extracted from three images (SCP, DCP, and full retina) with two segmentation masks (vessels and FAZ). Thus, in total, $(4 + 381 \times 2) \times 3 = 2298$ features were extracted for each patient.

Finally, in the classification step, data from these 2298 extracted features and demographic data for each patient were fed into four different machine learning classifiers—namely, Multiple Layer Perceptron (MLP), Random Forest (RF), Support Vector Machine, and eXtreme Gradient Boost (XGB)—to classify the OCTA images into six groups according to DR staging.[42–44]

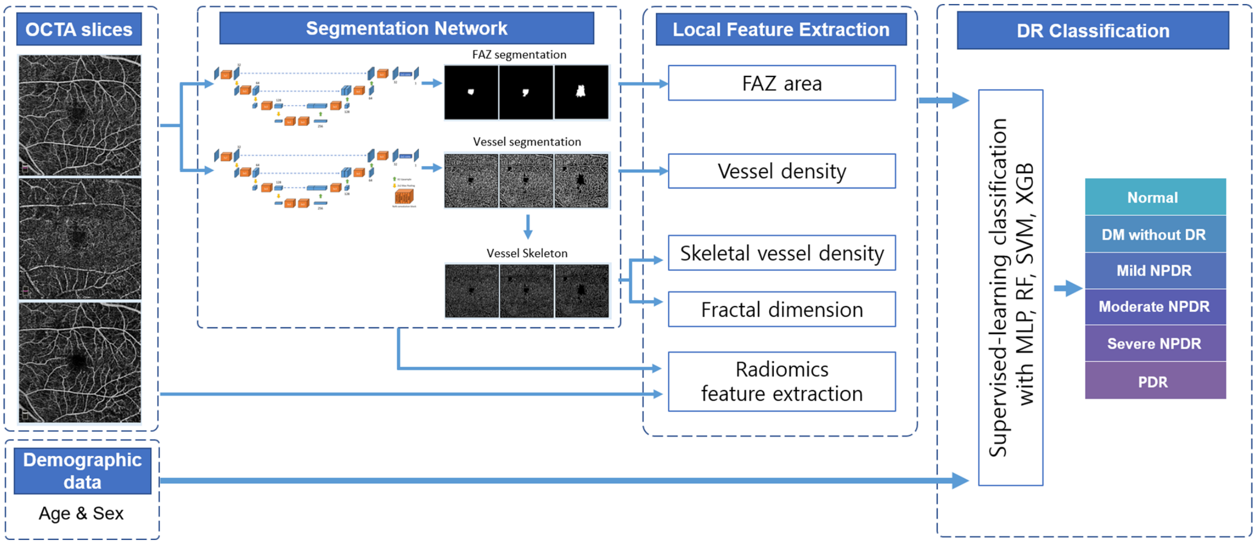

Figure 3. Machine learning–based classification networks for DR. FAZ and blood vessels were segmented from OCTA images using a U-Net model. The segmented FAZ and vessels were processed to extract four significant retinal features (i.e., the FAZ area, blood vessel density, skeletal vessel density, and fractal dimension). An additional 381 features were also extracted using the Pyradiomics toolbox. These features are supplied to four different machine learning based classifiers (i.e., MLP, RF, Support Vector Machine [SVM], and XGB) to classify the OCTA images into six groups according to DR staging.

Performance Evaluation of the Automated Classifiers

To obtain the final predictions for all data samples, we divided the data into four distinct subsets with an even class distribution. If not divisible by 4, the data were divided in such a way that the remainder was added to the fold one by one. Leaving one subset of the data for the test, a threefold cross-validation was conducted on each configuration, where the two subsets of the data were used for training and one for validation. All data were randomly divided at the patient level, and which data to use for training or testing was also randomly determined, but no data were used for both training and testing. The average metrics were derived from three test runs on the held-out data by comparing the predictions of the model with the gold standard determined by a retinal specialist using UWF FA.

Each of the automated classifiers based on a CNN and machine learning models was evaluated against the gold-standard UWF FA grades of the test set. For each classifier, the following metrics were evaluated: sensitivity, specificity, accuracy, F1 score, quadratic weighted κ, and area under the receiver operating characteristic (ROC) curve. These metrics were calculated based on the average value obtained by applying a stratified threefold cross-validation.

Performance Evaluation by Human Experts

Only 229 test data sets of 6 × 6 mm^2 scanned images were used to compare the performance of the algorithms with those of human experts. Each classification was conducted independently using OCTA images and the corresponding demographic data from the full data sets. Two retina specialists (GR and DP) independently classified the images by viewing them on a computer screen at a full image resolution. Only OCTA images and corresponding demographic data were provided for the classification task. These graders did not overlap with the retinal specialist who labeled the ground truth. The performance of each human expert was also evaluated and compared with the results from the automated classifiers.

Results

A total of 1118 eyes (254 healthy participants, 148 diagnosed with DM without DR, 108 with mild nonproliferative diabetic retinopathy [NPDR], 136 with moderate NPDR, 275 with severe NPDR, and 197 with proliferative diabetic retinopathy [PDR]) were recruited. After excluding the images with motion artifacts or an insufficient scan quality, 918 data sets of 3 × 3 mm^2 scanned images (219 healthy participants, 116 diagnosed with DM without DR, 81 with mild NPDR, 112 with moderate NPDR, 231 with severe NPDR, and 159 with PDR) and 917 data sets with 6 × 6 mm^2 scanned images (225 healthy participants, 104 diagnosed with DM without DR, 88 with mild NPDR, 105 with moderate NPDR, 234 with severe NPDR, and 161 with PDR) were obtained. The details of the data sets and splits are shown in Table 1, including the demographic characteristics of the study participants and the distribution of DR severity levels of the images.

Performance of CNN-Based Classifier

The performance of our CNN-based classifier was numerically superior using 6 × 6 mm^2 images than 3

Table 1. Numbers of Study Participants and OCTA Images Used for Deep Learning Model Training

Characteristic	Normal	DM Without DR	Mild NPDR	Moderate NPDR	Severe NPDR	PDR	Total
Participants, *n*	254	148	108	136	275	197	1118
Gender (male/female), *n*	85/169	83/65	73/35	69/67	167/108	115/82	592/526
Age, mean ± SD, y	62.5 ± 7.8	59.3 ± 13.5	61.0 ± 11.1	57.8 ± 11.0	58.2 ± 9.6	54.0 ± 10.3	58.8 ± 10.6
Laterality (right eye/left eye), *n*	125/129	84/64	61/47	65/71	125/150	94/103	554/564
OCTA images, *n*							
3 × 3 mm^2	219	116	81	112	231	159	918
Training and validation	164	87	61	84	171	118	685
Test	55	29	20	28	60	41	233
6 × 6 mm^2	225	104	88	105	234	161	917
Training and validation	169	78	66	79	175	121	688
Test	56	26	22	26	59	40	229

Table 2. Performance Results of Deep Learning CNN (ResNet 101) for Automated Classification of DR Severity

	Accuracy		Sensitivity		Specificity		F1 Score		QWK	
Characteristic	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3×3 mm^2 images										
Combined OCTA	0.684	0.005	0.617	0.001	0.934	0.002	0.634	0.001	0.856	0.017
Individual OCTA										
Superficial capillary plexus	0.681	0.019	0.619	0.017	0.934	0.004	0.635	0.022	0.840	0.014
Deep capillary plexus	0.651	0.007	0.590	0.019	0.927	0.001	0.609	0.031	0.847	0.005
Full-thickness retina	0.684	0.039	0.620	0.033	0.934	0.008	0.640	0.032	0.868	0.019
6×6 mm^2 images										
Combined OCTA	0.728	0.011	0.675	0.002	0.944	0.001	0.683	0.014	0.908	0.006
Individual OCTA										
Superficial capillary plexus	0.702	0.011	0.636	0.018	0.926	0.013	0.648	0.013	0.877	0.028
Deep capillary plexus	0.667	0.020	0.602	0.023	0.930	0.004	0.621	0.027	0.877	0.017
Full-thickness retina	0.721	0.025	0.657	0.019	0.941	0.004	0.681	0.032	0.882	0.012

Combined OCTA indicates that combined data of OCTA images (superficial capillary plexus, deep capillary plexus, and full-thickness retina layer) and demographic data (age and gender) were used as inputs. Individual OCTA indicates that individual layer of OCTA images and demographic data were used as inputs. QWK, quadratic weighted κ.

$\times$ 3 mm^2 images (Table 2). For the CNN-based classifier using combined data of 3×3 mm^2 OCTA images, the overall accuracy was 0.684, with an accuracy of 0.857 for detecting normal images, 0.883 for DM without DR, 0.910 for mild NPDR, 0.934 for moderate NPDR, 0.871 for severe NPDR, and 0.913 for PDR. The average sensitivity, specificity, F1 score, and quadratic weighted κ were 0.617, 0.934, 0.634, and 0.856, respectively. For the CNN-based classifier using combined data of 6×6 mm^2 OCTA images, the overall accuracy was measured to be 0.728; the accuracy was 0.849 for detecting normal images, 0.870 for DM without DR, 0.911 for mild NPDR, 0.956 for moderate NPDR, 0.920 for severe NPDR, and 0.949 for PDR. In addition, the average sensitivity, specificity, F1 score, and quadratic weighted κ were 0.675, 0.944, 0.683, and 0.908, respectively.

The performance of the CNN-based classifier was also evaluated using the SCP, DCP, and full-thickness retina layers separately (Table 2). The performance of the CNN using a separate OCTA layer of SCP or a full-thickness retina was nearly comparable to that of a CNN using combined data as input. However, the performances using the DCP layer were worse than those of the other layers.

The confusion matrices between the ground truth labels and the predictions of the proposed method are illustrated in Supplementary Figure S1, and the results for each of the three folds are listed in Supplementary Table S1. Figure 4 shows example CAM images in which the CNN correctly identified the DR staging using only OCTA images.

Performance of Machine Learning–Based Classifiers

The performances of the four machine learning classifiers are listed in Table 3. The average values obtained from the machine learning–based classifiers using local features extracted from the OCTA images were lower than the performance of the CNN classifier. Using 3×3 mm^2 OCTA images, XGB achieved the highest for three metrics (accuracy, specificity, and quadratic weighted κ) and MLP for the remaining two metrics (sensitivity and F1 score). Using 6×6 mm^2 OCTA images, RF achieved the highest for three metrics (accuracy, specificity, and F1 score) of the five other performance metrics relative to the other machine learning classifiers. In contrast to the CNN classifier, the machine learning–based classifier showed a better performance using 3×3 mm^2 rather than 6×6 mm^2 OCTA images.

The confusion matrices between the ground truth labels and the predictions of the four different machine learning classifiers are illustrated in Supplementary Figure S2, and the results for each of the three folds are listed in Supplementary Table S2.

Performance of Human Experts

A comparison of the performance of the CNN classifier, machine learning classifier, and human experts using 6×6 mm^2 OCTA images is shown in Table 4. As expected, the CNN classifier achieved the highest performance, and human experts achieved the lowest performance. The agreement between the two

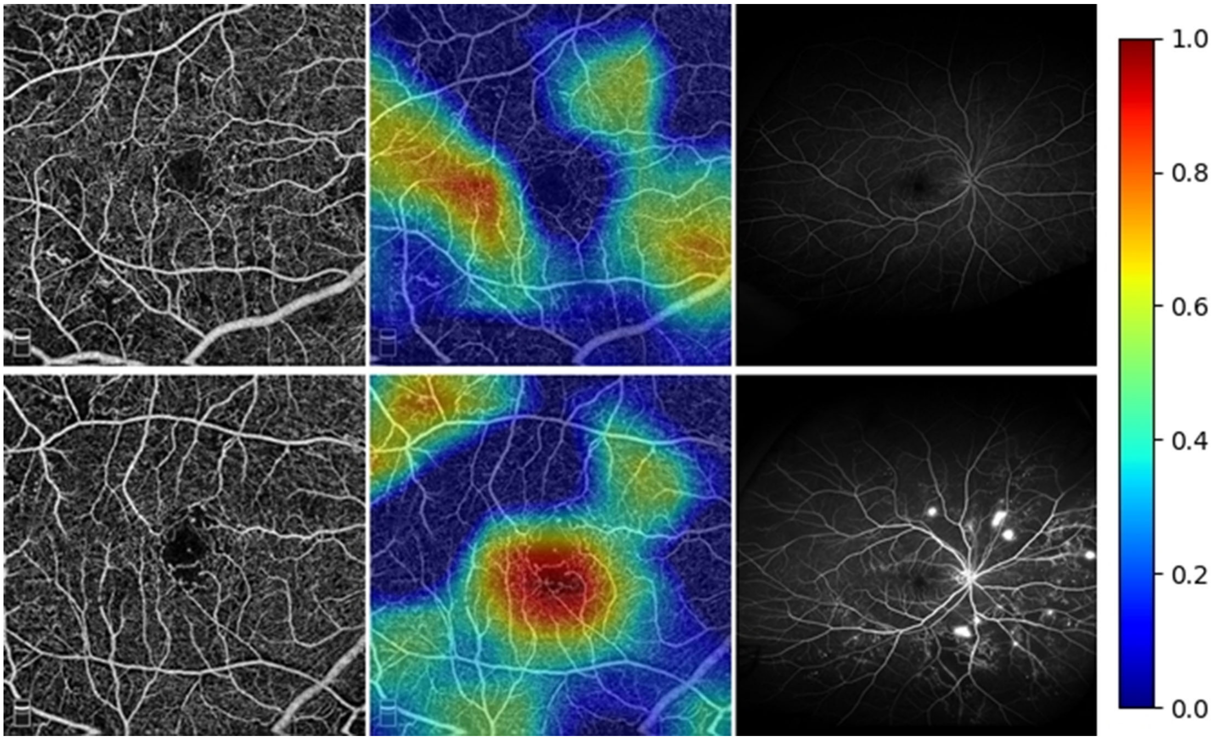

Figure 4. Representative examples where the CNN correctly identified the DR staging. For each of the two representative eyes, the original OCTA images (*left*), CAMs overlaid on the original images (*middle*), and corresponding ultra-widefield fluorescein angiograms (*right*) are shown. The heatmap scale for the CAMs is also shown with a signal range from zero (*purple*) to +1.00 (*brown*). The regions with high positive values (*red* to *brown*) in the CAM images were the regions the network used for decision-making. By contrast, the regions with nearly zero values (*blue* to *purple*) in the CAM images have no or negative influences on the classification. The *upper* images were derived from an eye with severe NPDR, and the *lower* images were derived from an eye with PDR. In cases similar to these examples, human grading of DR from OCTA images is difficult to perform accurately, even for trained graders. The *highlighted* regions in the CAM images corresponded well with pathologic regions such as where the density of blood vessels greatly changed or aneurysmal changes were observed.

experts was 0.463. A comparison of the ROC curve of the CNN classifier with machine learning classifiers and human experts is presented in Figure 5.

The confusion matrix between the ground truth labels and the predictions of the human experts is illustrated in Supplementary Figure S3, and the

Table 3. Performance Results of Four Different Machine Learning Classifiers (Multiple Layer Perceptron, Random Forest, Support Vector Machine, and eXtreme Gradient Boost) for Automated Classification of DR Severity

	Accuracy		Sensitivity		Specificity		F1 Score		QWK	
Characteristic	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3 × 3 mm^2 images										
Multiple Layer Perceptron	0.562	0.031	0.497	0.036	0.909	0.006	0.504	0.038	0.782	0.025
Random Forest	0.562	0.022	0.457	0.012	0.907	0.004	0.497	0.019	0.804	0.003
Support Vector Machine	0.468	0.038	0.423	0.033	0.893	0.007	0.425	0.033	0.731	0.035
eXtreme Gradient Boost	0.574	0.024	0.489	0.025	0.912	0.005	0.501	0.030	0.814	0.016
6 × 6 mm^2 images										
Multiple Layer Perceptron	0.502	0.037	0.430	0.031	0.897	0.008	0.432	0.036	0.781	0.003
Random Forest	0.531	0.011	0.408	0.014	0.898	0.003	0.435	0.023	0.785	0.001
Support Vector Machine	0.450	0.000	0.397	0.002	0.887	0.001	0.398	0.004	0.735	0.002
eXtreme Gradient Boost	0.508	0.018	0.417	0.013	0.897	0.004	0.425	0.014	0.802	0.006

The combined data of OCTA images (superficial capillary plexus, deep capillary plexus, and full-thickness retina layer) and demographic data (age and gender) were used as inputs for development of the machine learning classifiers.

Table 4. Comparison of the Performance Results Between the CNN, Machine Learning Classifier, and Human Experts for DR Classification From OCTA Images

	Accuracy		Sensitivity		Specificity		F1 Score		QWK	
Characteristic	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CNN classifier	0.728	0.011	0.675	0.002	0.944	0.001	0.683	0.014	0.908	0.006
Machine learning classifier (RF)	0.531	0.011	0.408	0.014	0.898	0.003	0.435	0.023	0.785	0.001
Human experts (average)	0.330	0.020	0.320	0.023	0.869	0.004	0.337	0.022	0.713	0.020

Only 229 test data sets of 6 × 6 mm^2 scan images were used to compare the performance results between the convolutional neural network, machine learning classifier, and human experts. As a representative of the machine learning classifier, the results of RF, the most accurate machine learning classifier used in this study, are described in the table. For the performance results of human experts, the average values of the performance results of the two experts are also described.

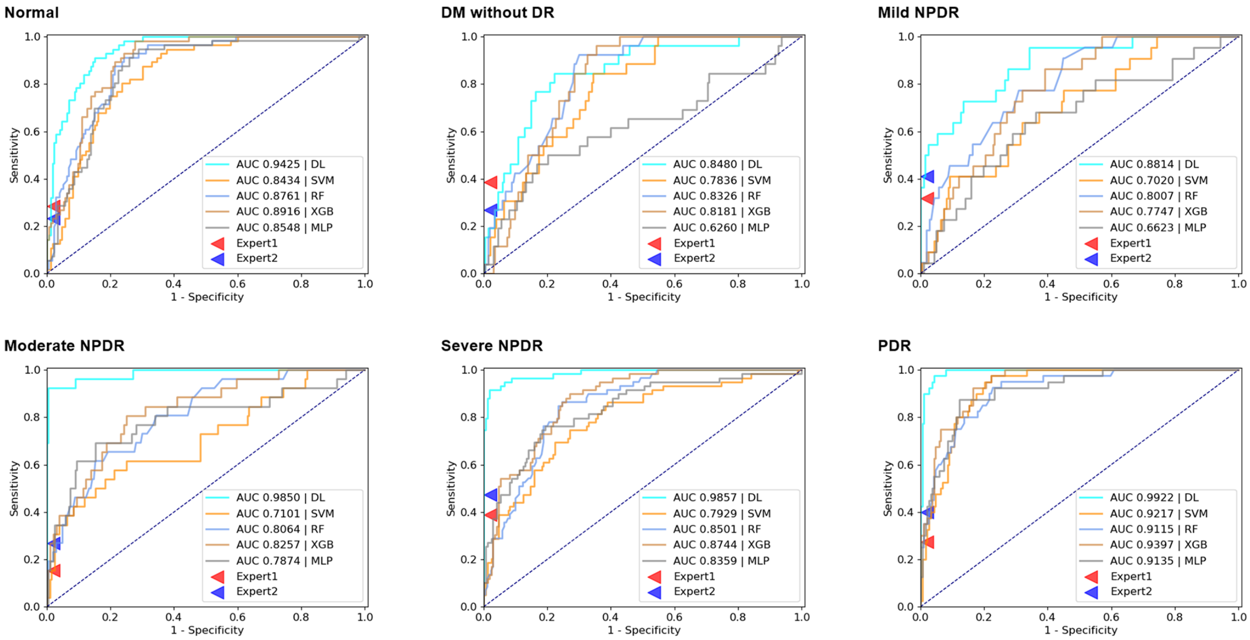

Figure 5. ROC curves illustrating classification performances for the prediction of DR staging. The *dotted line* represents the trade-off resulting from random chance, and the *solid curved lines* represent the trade-off of each automated classifier. The performances of retinal experts are plotted as *red* and *blue triangles*.

results for each expert are listed in Supplementary Table S3.

Discussion

In this study, we proposed an end-to-end CNN architecture using OCTA for automated DR classification. As expected, the end-to-end CNN classifier outperformed the machine learning classifiers, which used 2298 extracted local features of each OCTA image to classify the images into six groups according to DR staging. Radiomics is a systematic approach for studying latent information in medical imaging for improved accuracy. Among them, PyRadiomics is the most widely reported radiomics tool in the literature, and it contains thousands of handcrafted formulas designed to extract the distribution or texture information from medical images.[41] Although these feature-based methods achieved lower classification performance than the CNN method, they can find out which features have a major impact on the classification through the random forest or L1 optimization. By contrast, since the CNN method operates end-to-end without explicitly extracting features, it is difficult to know which features have a major influence on classification. However, the parameters in a CNN can be updated during backpropagation from the feature extraction perspective, allowing the extraction of a larger number of features that are associated with the target outcome. Because OCTA contains more unlabeled information, a fully automated CNN algorithm can process heterogeneous images quickly for an accurate and objective DR classification, potentially alleviating the requirement for a resource-intensive manual analysis and thus guiding high-risk patients for further treatment.

Also, the activation map allowed us to identify the areas in which the network was used for decision-making. By visualizing the CAM, we may identify informative image patterns or features that are useful for DR staging. However, the interpretation of these results warrants additional scrutiny because recent studies emphasized that many popular saliency maps used to interpret CNN trained on medical imaging did not meet several key criteria for utility and robustness, highlighting the need for additional validation before clinical application.[45–47] For the alternative technique, a computer-aided diagnosis system that utilizes the complementary information from CNN-based and feature-based methods will need to be further developed. Also, qualitative analysis of the latest techniques to better obtain the activation map will be required.[45]

When comparing the performance of the CNN algorithm according to the input image size, OCTA images covering a larger 6 × 6 mm^2 scanned area provided a higher performance than images covering a smaller scanned area. These results strongly support previous suggestions that wider fields of view may be more desirable for early detection and monitoring of disease progression.[48–50] Meanwhile, the ML classifier showed a better performance using 3 × 3 mm^2 OCTA images, which is a completely opposite result from that of the CNN classifier. We suspect that the cause of this discrepancy is a problem in the process of extracting handcrafted features through multiple steps. Motion artifacts and distorted weak signal regions are observed more frequently in widefield OCTA, particularly in the periphery. Moreover, large SCP vessels observed in the 6 × 6 mm^2 OCTA images may substantially contribute to the assessment of DR owing to their large diameters, which is not observed in 3 × 3 mm^2 OCTA images. Although a decreased capillary perfusion and an increased capillary dropout area have been reported to be associated with worsening DR severity, larger retinal arteriolar and venular calibers are also known to increase with the DR progression.[51–54] Because machine learning classifiers use quantitative parameters of OCTA images to classify the DR severity, the scan size can affect the results, particularly during the segmentation or feature extraction stage.

Interestingly, we also observed that the CNN algorithm for DR classification achieved poor results when using the DCP layer in comparison to other OCTA layers. Because the pathology in DR is hypothesized to preferentially involve a more vulnerable DCP, the results may appear to be contrary to common knowledge.[55] There are several potential explanations. Because the CNN algorithm is trained and tested based on FA images, which only visualize the superficial retinal vessels, it is perhaps not surprising that the CNN appears to perform better using SCP images than DCP images.[56] In addition, images of DCP layers may have been affected by projection artifacts caused by shadows from superficial blood flow projected onto deeper layers, resulting in an erroneous perception of flow. Because the deeper layers are more susceptible to projection artifacts and signal attenuation, this can potentially explain the greater variation in the interpretation of OCTA images in the DCP.[57] Similar to the results, several previous studies have also suggested that SCP continues to have a greater diagnostic value even after the DCP image quality has been improved through the removal of decorrelation tail projection artifacts.[58,59]

Since the wide use of CNN methods for image classification problems, several methods for the automated classification of DR severity have been proposed.[10–21,60] Most of these methods are based on fundus photographs. Ghosh et al.[19] proposed a CNN-based method to classify fundus photography into five classes (no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR) and achieved an overall accuracy of 85%. Owing to the restricted data set size compared to the extremely large fundus photography data set used in the previous fundus photography–based networks, many fewer studies have focused on the CNN algorithm applied to OCT and OCTA. However, OCT and OCTA have advantages over fundus photography in that they provide more instructive information on the structure and vasculature of the retina. Zang et al.[11] applied deep learning approaches to automated DR classification based on OCT and OCTA data and achieved an overall accuracy of 71% for the classification into four classes (no DR, mild and moderate NPDR, severe NPDR, and PDR), which is a slightly lower accuracy compared to fundus photography–based DR classifications. The authors pointed out that this is due to the relatively small data sets (approximately 1/100 of that of previous studies using fundus photography data sets) and the use of an algorithm trained with classification based on fundus photography, which is a considerably different modality from OCT/OCTA. Although multiple studies have examined various artificial intelligence-based approaches to the classification of DR, we are unaware of any algorithm trained with classification based on UWF FA. In previous studies, the grading system of DR was based on a fundus photograph examination, making it prone to oversight of subtle fundus details, leading to examiner errors. In addition, alterations of the microcirculation in the peripheral retina were not observed upon a fundus examination. A recent study revealed that 17% of retinal neovascularization lies anterior to the border of seven conventional standard fields, suggesting that UWF FA allows for a more appropriate staging of DR.[24]

Although we reported a comparable performance in this study, as a notable limitation, the number of patients employed is still relatively small. However, the number of patients in this study is comparable to that in others employing OCTA,[10,11,20,21,40] considering that this technology is still not ubiquitous in ophthalmology practices. We used training and testing OCTA data from only a single center, without generalizability testing using external data sets. Further studies conducting robust prospective external validation tests are required. Also, it is necessary to compare performance for DR classification between ResNet 101 and other CNN architectures (e.g., DenseNet, EfficientNet, or Inception v3). However, this study supports an important first step in end-to-end deep learning models for DR classification using OCTA images. As a strength of this study, the ground truth for the classification of DR stages is based on the UWF FA. Although OCTA has several clinical advantages over FA, its role in the clinical decision-making process is still limited. Using the CNN algorithm, we can classify the DR severity in an automated fashion by taking advantage of both UWF FA and OCTA.

In this study, we introduced a fully automated deep CNN DR classification method using OCTA images. Although OCTA is rapidly adapted to the new modality in a clinical routine, the interpretation of OCTA data remains limited. If the proposed automated DR classification framework using OCTA can provide a similar level of diagnostic value as other modalities, the number of procedures an individual would require for an accurate diagnosis would be reduced, ultimately lowering both the clinical burden and the health care costs. This system is expected to drastically reduce, on a clinical basis, the rate of vision loss attributed to DR; improve clinical management; and create a novel diagnostic workflow for disease detection and referral. For a proper clinical application of our method, further testing and optimization of the sensitivity metrics, such as genetic factors, hemoglobin A1C, duration of diabetes, and other clinical data, may be required to ensure a minimum false-negative rate. Combining the data from various imaging modalities, such as fundus photography or FA, can reinforce the performance value and thereby further improve the accuracy. Future work should include extending the algorithm to a larger number of participants, even including images with macular edema, artifacts, or low quality, to make it more generalizable in a practical manner.

Acknowledgments

Supported by a 2020 Yeungnam University research grant (to MS) and the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2019R1C1C1008727; to SHP).

Disclosure: **G. Ryu,** None; **K. Lee,** None; **D. Park,** None; **I. Kim,** None; **S.H. Park,** None; **M. Sagong,** None

[*] GR and KL contributed equally as first authors.

References

1. Elman MJ, Aiello LP, Beck RW, et al. Randomized trial evaluating ranibizumab plus prompt or deferred laser or triamcinolone plus prompt laser for diabetic macular edema. *Ophthalmology*. 2010;117(6):1064–1077.e35.
2. Massin P, Bandello F, Garweg JG, et al. Safety and efficacy of ranibizumab in diabetic macular edema (RESOLVE study): a 12-month, randomized, controlled, double-masked, multicenter phase II study. *Diabetes Care*. 2010;33(11):2399–2405.
3. Michaelides M, Kaines A, Hamilton RD, et al. A prospective randomized trial of intravitreal bevacizumab or laser therapy in the management of diabetic macular edema (BOLT study): 12-month data: report 2. *Ophthalmology*. 2010;117(6):1078–1086.e2.
4. Mitchell P, Bandello F, Schmidt-Erfurth U, et al. The RESTORE study: ranibizumab monotherapy or combined with laser versus laser monotherapy for diabetic macular edema. *Ophthalmology*. 2011;118(4):615–625.
5. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. *Expert Rev Ophthalmol*. 2012;7(5):417–439.
6. Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care*. 2017;40(3):412–418.
7. Sellahewa L, Simpson C, Maharajan P, Duffy J, Idris I. Grader agreement, and sensitivity and

specificity of digital photography in a community optometry-based diabetic eye screening program. *Clin Ophthalmol*. 2014;8:1345.

8. Ruamviboonsuk P, Wongcumchang N, Surawongsin P, Panyawatananukul E, Tiensuwan M. Screening for diabetic retinopathy in rural area using single-field, digital fundus images. *J Med Assoc Thai*. 2005;88(2):176–180.
9. Zhou Y, Lu S. Discovering abnormal patches and transformations of diabetics retinopathy in big fundus collections. *Computer Science & Information Technology (CS & IT)*. 2017;7(1):195–206.
10. Heisler M, Karst S, Lo J, et al. Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography. *Transl Vis Sci Technol*. 2020;9(2):20.
11. Zang P, Gao L, Hormel TT, et al. DcardNet: diabetic retinopathy classification at multiple levels based on structural and angiographic optical coherence tomography. *IEEE Trans Biomed Eng*. 2021;68(6):1859–1870.
12. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health*. 2020;2(5):e240–e249.
13. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 2020;34(3):451–460.
14. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2(1):1–9.
15. Akram MU, Khalid S, Tariq A, Khan SA, Azam F. Detection and classification of retinal lesions for grading of diabetic retinopathy. *Comput Biol Med*. 2014;45:161–171.
16. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200–5206.
17. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
18. Alam M, Zhang Y, Lim JI, Chan RV, Yang M, Yao X. Quantitative optical coherence tomography angiography features for objective classification and staging of diabetic retinopathy. *Retina*. 2020;40(2):322–332.
19. Ghosh R, Ghosh K, Maitra S. Automatic detection and classification of diabetic retinopathy stages using CNN. In *Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*. Noida, India, February 2–3, 2017;550–554.
20. Sandhu HS, Eladawi N, Elmogy M, et al. Automated diabetic retinopathy detection using optical coherence tomography angiography: a pilot study. *Br J Ophthalmol*. 2018;102(11):1564–1569.
21. Sandhu HS, Elmogy M, Sharafeldeen AT, et al. Automated diagnosis of diabetic retinopathy using clinical biomarkers, optical coherence tomography, and optical coherence tomography angiography. *Am J Ophthalmol*. 2020;216:201–206.
22. Silva PS, Cavallerano JD, Haddad NMN, et al. Peripheral lesions identified on ultrawide field imaging predict increased risk of diabetic retinopathy progression over 4 years. *Ophthalmology*. 2015;122(5):949–956.
23. Silva PS, Cruz AJD, Ledesma MG, et al. Diabetic retinopathy severity and peripheral lesions are associated with nonperfusion on ultrawide field angiography. *Ophthalmology*. 2015;122(12):2465–2472.
24. Wessel MM, Aaker GD, Parlitsis G, Cho M, D'Amico DJ, Kiss S. Ultra–wide-field angiography improves the detection and classification of diabetic retinopathy. *Retina*. 2012;32(4):785–791.
25. Price LD, Au S, Chong NV. Optomap ultrawide field imaging identifies additional retinal abnormalities in patients with diabetic retinopathy. *Clin Ophthalmol*. 2015;9:527.
26. Kwiterovich KA, Maguire MG, Murphy RP, et al. Frequency of adverse systemic reactions after fluorescein angiography: results of a prospective study. *Ophthalmology*. 1991;98(7):1139–1142.
27. Spaide RF, Fujimoto JG, Waheed NK, Sadda SR, Staurenghi G. Optical coherence tomography angiography. *Prog Retin Eye Res*. 2018;64:1–55.
28. Ang M, Tan AC, Cheung CMG, et al. Optical coherence tomography angiography: a review of current and future clinical applications. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(2):237–245.
29. Glacet-Bernard A, Miere A, Houmane B, Tilleul J, Souied E. Nonperfusion assessment in retinal vein occlusion: comparison between ultra-widefield fluorescein angiography and widefield optical coherence tomography angiography. *Retina*. 2021;41(6):1202–1209.
30. Cui Y, Zhu Y, Wang JC, et al. Comparison of widefield swept-source optical coherence tomography angiography with ultra-widefield colour fundus photography and fluorescein angiography for detection of lesions in diabetic retinopathy. *Br J Ophthalmol*. 2021;105(4):577–581.

31. Khalid H, Schwartz R, Nicholson L, et al. Widefield optical coherence tomography angiography for early detection and objective evaluation of proliferative diabetic retinopathy. *Br J Ophthalmol*. 2021;105(1):118–123.
32. Ryu G, Park D, Lim J, van Hemert J, Sagong M. Macular microvascular changes and their correlation with peripheral nonperfusion in branch retinal vein occlusion. *Am J Ophthalmol*. 2021;225:57–68.
33. Barekatain K, Sharma S, Ehlers JP, et al. Swept-source optical coherence tomography angiography parameters correlate with leakage and ischemic indices from ultra-widefield fluorescein angiography in diabetic retinopathy. *Invest Ophthalmol Vis Sci*. 2020;61(7):4100–4100.
34. Wilkinson C, Ferris FL, III, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110(9):1677–1682.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, United States: IEEE; 2016:770–778.
36. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, United States: IEEE; 2016:2921–2929.
37. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Miami, United States: IEEE; 2009;248–255.
38. Kingma DP, Ba J. Adam: A method for stochastic optimization. *Anon. International Conference on Learning Representations.* SanDego: ICLR; 2015.
39. Ronneberger O, Fischer P, Brox T. In: Navab N, Hornegger J, Wells W, Frangi A, eds. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol 9351.* Cham, Switzerland: Springer; 2015:234–241.
40. Eladawi N, Elmogy M, Ghazal M, et al. Early signs detection of diabetic retinopathy using optical coherence tomography angiography scans based on 3D multi-path convolutional neural network. *2019 IEEE International Conference on Image Processing (ICIP).* Taipei, Taiwan; September 22–25, 2019:1390–1394.
41. Van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017; 77(21):e104–e107.
42. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
43. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273–297.
44. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, United States: Association for Computing Machiner; 2016:785–794.
45. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology*. 2021;3(6):e200267.
46. Singh A, Jothi Balaji J, Rasheed MA, Jayakumar V, Raman R, Lakshminarayanan V. Quantitative and qualitative evaluation of explainable deep learning methods for ophthalmic diagnosis. arXiv e-prints. 2020:arXiv-2009.
47. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med Phys*. 2016;43(12):6654–6666.
48. Alibhai AY, De Pretto LR, Moult EM, et al. Quantification of retinal capillary nonperfusion in diabetics using wide-field optical coherence tomography angiography. *Retina*. 2020;40(3):412–420.
49. Wang F, Saraf SS, Zhang Q, Wang RK, Rezaei KA. Ultra-widefield protocol enhances automated classification of diabetic retinopathy severity with OCT angiography. *Ophthalmol Retina*. 2020;4(4):415–424.
50. Kim K, You JI, Park JR, Kim ES, Oh W-Y, Yu S-Y. Quantification of retinal microvascular parameters by severity of diabetic retinopathy using wide-field swept-source optical coherence tomography angiography. *Graefes Arch Clin Exp Ophthalmol*. 2021;259(8):2103–2111.
51. Sun C, Wang JJ, Mackey DA, Wong TY. Retinal vascular caliber: systemic, environmental, and genetic associations. *Surv Ophthalmol*. 2009; 54(1):74–95.
52. Ikram MK, Ong YT, Cheung CY, Wong TY. Retinal vascular caliber measurements: clinical significance, current knowledge and future perspectives. *Ophthalmologica*. 2013;229(3):125–136.
53. Grunwald JE, DuPont J, Riva CE. Retinal haemodynamics in patients with early diabetes mellitus. *Br J Ophthalmol*. 1996;80(4):327–331.

54. Tan B, Chua J, Lin E, et al. Quantitative microvascular analysis with wide-field optical coherence tomography angiography in eyes with diabetic retinopathy. *JAMA Netw Open*. 2020; 3(1):e1919469.
55. Zhu TP, Li EH, Li JY, et al. Comparison of projection-resolved optical coherence tomography angiography-based metrics for the early detection of retinal microvascular impairments in diabetes mellitus. *Retina*. 2020;40(9):1783–1792.
56. Spaide RF, Klancnik JM, Cooney MJ. Retinal vascular layers imaged by fluorescein angiography and optical coherence tomography angiography. *JAMA Ophthalmol*. 2015;133(1):45–50.
57. Spaide RF, Fujimoto JG, Waheed NK. Image artifacts in optical coherence angiography. *Retina*. 2015;35(11):2163.
58. Durbin MK, An L, Shemonski ND, et al. Quantification of retinal microvascular density in optical coherence tomographic angiography images in diabetic retinopathy. *JAMA Ophthalmol*. 2017;135(4):370–376.
59. Ong JX, Kwan CC, Cicinelli MV, Fawzi AA. Superficial capillary perfusion on optical coherence tomography angiography differentiates moderate and severe nonproliferative diabetic retinopathy. *PLoS One*. 2020;15(10): e0240064.
60. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7):962–969.